SYNFloodPDA

Taulant, Amir, Jatin, Carter

2025-03-12

Import Libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

Import Balanced dataset

```
# Import syn dataset
syn_dt <- read.csv("Balanced-SYN-V2.csv")

# Explore the dataset
str(syn_dt)
```

```
## 'data.frame':    10000 obs. of  88 variables:
##  $ Unnamed..0                : int  7372 13359 131547 4879 9939 10158 1254 477265 1624 1713 ...
##  $ Flow.ID                   : chr  "192.168.50.6-4.2.2.4-54280-53-17" "192.168.50.6-8.8.8.8-58920-5
##  $ Source.IP                 : chr  "192.168.50.6" "192.168.50.6" "172.217.10.67" "192.168.50.9" ..
##  $ Source.Port               : int  54280 58920 443 36470 443 62429 57363 62138 50764 465 ...
##  $ Destination.IP            : chr  "4.2.2.4" "8.8.8.8" "192.168.50.9" "172.217.9.226" ...
##  $ Destination.Port          : int  53 53 43594 443 57078 80 80 53 443 37826 ...
##  $ Protocol                  : int  17 17 6 6 6 6 6 17 6 6 ...
##  $ Timestamp                 : chr  "2018-11-03 15:12:32.442181" "2018-11-03 13:37:04.525418" "2018
##  $ Flow.Duration             : int  25040 47213 110 118548880 147 7345102 707372 21190 0 98 ...
##  $ Total.Fwd.Packets         : int  2 2 1 40 1 2 7 2 2 1 ...
##  $ Total.Backward.Packets    : int  2 2 2 46 2 8 2 2 0 3 ...
##  $ Total.Length.of.Fwd.Packets: num  84 86 0 4022 6 ...
##  $ Total.Length.of.Bwd.Packets: num  140 118 12 12880 12 ...
##  $ Fwd.Packet.Length.Max     : num  42 43 0 605 6 0 597 34 24 0 ...
##  $ Fwd.Packet.Length.Min     : num  42 43 0 0 6 0 0 34 6 0 ...
##  $ Fwd.Packet.Length.Mean    : num  42 43 0 101 6 ...
```

```
##  $ Fwd.Packet.Length.Std      : num  0 0 0 181 0 ...
##  $ Bwd.Packet.Length.Max      : num  70 59 6 1418 6 ...
##  $ Bwd.Packet.Length.Min      : num  70 59 6 0 6 0 0 282 0 0 ...
##  $ Bwd.Packet.Length.Mean     : num  70 59 6 280 6 ...
##  $ Bwd.Packet.Length.Std      : num  0 0 0 462 0 ...
##  $ Flow.Bytes.s               : num  8946 4321 109091 143 122449 ...
##  $ Flow.Packets.s             : num  1.60e+02 8.47e+01 2.73e+04 7.25e-01 2.04e+04 ...
##  $ Flow.IAT.Mean              : num  8.35e+03 1.57e+04 5.50e+01 1.39e+06 7.35e+01 ...
##  $ Flow.IAT.Std               : num  1.45e+04 2.73e+04 7.35e+01 8.94e+06 9.97e+01 ...
##  $ Flow.IAT.Max               : num  25036 47208 107 58876783 144 ...
##  $ Flow.IAT.Min               : num  2 2 3 1 3 1 3 3 0 1 ...
##  $ Fwd.IAT.Total              : num  2.00 2.00 0.00 1.19e+08 0.00 ...
##  $ Fwd.IAT.Mean               : num  2 2 0 3039715 0 ...
##  $ Fwd.IAT.Std                : num  0 0 0 13095439 0 ...
##  $ Fwd.IAT.Max                : num  2 2 0 58917906 0 ...
##  $ Fwd.IAT.Min                : num  2 2 0 1 0 3 3 3 0 0 ...
##  $ Bwd.IAT.Total              : num  2.00 3.00 3.00 1.19e+08 3.00 ...
##  $ Bwd.IAT.Mean               : num  2 3 3 2633510 3 ...
##  $ Bwd.IAT.Std                : num  0 0 0 12223952 0 ...
##  $ Bwd.IAT.Max                : num  2 3 3 58917838 3 ...
##  $ Bwd.IAT.Min                : num  2 3 3 1 3 1 3 3 0 1 ...
##  $ Fwd.PSH.Flags              : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Bwd.PSH.Flags              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fwd.URG.Flags              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Bwd.URG.Flags              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fwd.Header.Length          : int  64 40 32 1296 20 64 164 64 40 32 ...
##  $ Bwd.Header.Length          : int  40 40 40 1488 40 256 64 64 0 96 ...
##  $ Fwd.Packets.s              : num  79.872 42.361 9090.909 0.337 6802.721 ...
##  $ Bwd.Packets.s              : num  7.99e+01 4.24e+01 1.82e+04 3.88e-01 1.36e+04 ...
##  $ Min.Packet.Length          : num  42 43 0 0 6 0 0 34 6 0 ...
##  $ Max.Packet.Length          : num  70 59 6 1418 6 ...
##  $ Packet.Length.Mean         : num  53.2 49.4 3 194.3 6 ...
##  $ Packet.Length.Std          : num  15.34 8.76 3.46 367.1 0 ...
##  $ Packet.Length.Variance     : num  235.2 76.8 12 134765.5 0 ...
##  $ FIN.Flag.Count             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SYN.Flag.Count             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RST.Flag.Count             : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ PSH.Flag.Count             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ACK.Flag.Count             : int  0 0 0 1 0 1 1 0 0 0 ...
##  $ URG.Flag.Count             : int  0 0 1 0 1 0 0 0 1 1 ...
##  $ CWE.Flag.Count             : int  0 0 1 0 1 0 0 0 0 1 ...
##  $ ECE.Flag.Count             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Down.Up.Ratio              : num  1 1 2 1 2 4 0 1 0 3 ...
##  $ Average.Packet.Size        : num  66.5 61.8 4 196.5 8 ...
##  $ Avg.Fwd.Segment.Size       : num  42 43 0 101 6 ...
##  $ Avg.Bwd.Segment.Size       : num  70 59 6 280 6 ...
##  $ Fwd.Header.Length.1        : int  64 40 32 1296 20 64 164 64 40 32 ...
##  $ Fwd.Avg.Bytes.Bulk         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fwd.Avg.Packets.Bulk       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fwd.Avg.Bulk.Rate          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Bwd.Avg.Bytes.Bulk         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Bwd.Avg.Packets.Bulk       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Bwd.Avg.Bulk.Rate          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Subflow.Fwd.Packets        : int  2 2 1 40 1 2 7 2 2 1 ...
```

```
##  $ Subflow.Fwd.Bytes          : int  84 86 0 4022 6 0 1212 68 30 0 ...
##  $ Subflow.Bwd.Packets        : int  2 2 2 46 2 8 2 2 0 3 ...
##  $ Subflow.Bwd.Bytes          : int  140 118 12 12880 12 0 0 564 0 74 ...
##  $ Init_Win_bytes_forward     : int  -1 -1 244 29200 250 8192 8192 -1 297 244 ...
##  $ Init_Win_bytes_backward    : int  -1 -1 0 253 258 29200 29200 -1 -1 245 ...
##  $ act_data_pkt_fwd           : int  1 1 0 20 0 0 5 1 1 0 ...
##  $ min_seg_size_forward       : int  32 20 32 32 20 32 20 32 20 32 ...
##  $ Active.Mean                : num  0 0 0 639677 0 ...
##  $ Active.Std                 : num  0 0 0 785271 0 ...
##  $ Active.Max                 : num  0 0 0 1194947 0 ...
##  $ Active.Min                 : num  0 0 0 84406 0 ...
##  $ Idle.Mean                  : num  0 0 0 58614127 0 ...
##  $ Idle.Std                   : num  0 0 0 371452 0 ...
##  $ Idle.Max                   : num  0 0 0 58876783 0 ...
##  $ Idle.Min                   : num  0 0 0 58351470 0 ...
##  $ SimillarHTTP               : chr  "0" "0" "0" "0" ...
##  $ Inbound                    : int  0 0 1 0 1 0 0 0 0 1 ...
##  $ Label                      : chr  "BENIGN" "BENIGN" "BENIGN" "BENIGN" ...
```

Data Preprocessing

```
# Select relevant variables with dyplr library

syn_selected_dt <- syn_dt %>% select('SYN.Flag.Count',
                                     'Total.Fwd.Packets',
                                     'Total.Backward.Packets',
                                     'Flow.Duration',
                                     'Flow.Packets.s',
                                     'Flow.Bytes.s',
                                     'Fwd.Packet.Length.Mean',
                                     'Bwd.Packet.Length.Mean',
                                     'Bwd.IAT.Mean',
                                     'ACK.Flag.Count',
                                     'Active.Mean',
                                     'Label',
                                     'Inbound')
```

```
# Use summary function to explore selected dataset
summary(syn_selected_dt)
```

```
##  SYN.Flag.Count   Total.Fwd.Packets  Total.Backward.Packets Flow.Duration
##  Min.   :0.0000   Min.   :   1.000   Min.   :   0.000       Min.   :        0
##  1st Qu.:0.0000   1st Qu.:   2.000   1st Qu.:   0.000       1st Qu.:        1
##  Median :0.0000   Median :   2.000   Median :   2.000       Median :      104
##  Mean   :0.0032   Mean   :   6.678   Mean   :   6.242       Mean   :  9825612
##  3rd Qu.:0.0000   3rd Qu.:   2.000   3rd Qu.:   2.000       3rd Qu.:    33506
##  Max.   :1.0000   Max.   :3890.000   Max.   :6706.000       Max.   :119991155
##
##  Flow.Packets.s       Flow.Bytes.s      Fwd.Packet.Length.Mean
##  Min.   :      0.1   Min.   :       0   Min.   :   0.00
##  1st Qu.:    151.2   1st Qu.:    4098   1st Qu.:   6.00
##  Median :  37735.8   Median :  220680   Median :   6.00
##  Mean   :      Inf   Mean   :     Inf   Mean   :  23.93
##  3rd Qu.:2000000.0   3rd Qu.:12000000   3rd Qu.:  29.00
```

```
##  Max.   :      Inf   Max.   :     Inf   Max.   :1797.62
##                     NA's   :4
##  Bwd.Packet.Length.Mean  Bwd.IAT.Mean      ACK.Flag.Count    Active.Mean
##  Min.   :   0.00        Min.   :       0   Min.   :0.0000   Min.   :       0
##  1st Qu.:   0.00        1st Qu.:       0   1st Qu.:0.0000   1st Qu.:       0
##  Median :   6.00        Median :       1   Median :1.0000   Median :       0
##  Mean   :  52.34        Mean   :  398482   Mean   :0.5998   Mean   :   65872
##  3rd Qu.:   6.00        3rd Qu.:       3   3rd Qu.:1.0000   3rd Qu.:       0
##  Max.   :1792.35        Max.   :34909930   Max.   :1.0000   Max.   :10073804
##
##     Label              Inbound
##  Length:10000       Min.   :0.0000
##  Class :character   1st Qu.:0.0000
##  Mode  :character   Median :1.0000
##                     Mean   :0.6053
##                     3rd Qu.:1.0000
##                     Max.   :1.0000
##
```

Data Cleaning

```
# Removing NA values
syn_selected_dt <- na.omit(syn_selected_dt)
```

```
# Removing infinitty values
syn_selected_dt[syn_selected_dt == "Inf"] <- NA
syn_selected_dt <- na.omit(syn_selected_dt)
```

```
str(syn_selected_dt)
```

```
## 'data.frame':    9604 obs. of  13 variables:
##  $ SYN.Flag.Count        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Total.Fwd.Packets     : int  2 2 1 40 1 2 7 2 1 2 ...
##  $ Total.Backward.Packets: int  2 2 2 46 2 8 2 2 3 0 ...
##  $ Flow.Duration         : int  25040 47213 110 118548880 147 7345102 707372 21190 98 1 ...
##  $ Flow.Packets.s        : num  1.60e+02 8.47e+01 2.73e+04 7.25e-01 2.04e+04 ...
##  $ Flow.Bytes.s          : num  8946 4321 109091 143 122449 ...
##  $ Fwd.Packet.Length.Mean: num  42 43 0 101 6 ...
##  $ Bwd.Packet.Length.Mean: num  70 59 6 280 6 ...
##  $ Bwd.IAT.Mean          : num  2 3 3 2633510 3 ...
##  $ ACK.Flag.Count        : int  0 0 0 1 0 1 1 0 0 0 ...
##  $ Active.Mean           : num  0 0 0 639677 0 ...
##  $ Label                 : chr  "BENIGN" "BENIGN" "BENIGN" "BENIGN" ...
##  $ Inbound               : int  0 0 1 0 1 0 0 0 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:392] 9 29 148 363 648 649 712 739 811 840 ...
##   ..- attr(*, "names")= chr [1:392] "9" "29" "148" "363" ...
```

```
table(syn_selected_dt$Label)
```

```
##
## BENIGN    Syn
##   4949   4655
```

```
str(syn_selected_dt$Flow.Duration)
```

```
##  int [1:9604] 25040 47213 110 118548880 147 7345102 707372 21190 98 1 ...
```

```r
# Step 1: Convert 'BENIGN' to "0", everything else to "1"
syn_selected_dt$Label[syn_selected_dt$Label == "BENIGN"] <- "0"
syn_selected_dt$Label[syn_selected_dt$Label != "0"] <- "1"

# Step 2: Convert to numeric
syn_selected_dt$Label <- as.numeric(syn_selected_dt$Label)

# Step 3: Convert to factor with levels 0 and 1
syn_selected_dt$Label <- factor(syn_selected_dt$Label, levels = c(0, 1))
```

```r
summary(syn_selected_dt)
```

```
##  SYN.Flag.Count     Total.Fwd.Packets  Total.Backward.Packets
##  Min.   :0.000000   Min.   :   1.000   Min.   :   0.0
##  1st Qu.:0.000000   1st Qu.:   2.000   1st Qu.:   0.0
##  Median :0.000000   Median :   2.000   Median :   2.0
##  Mean   :0.003332   Mean   :   6.871   Mean   :   6.5
##  3rd Qu.:0.000000   3rd Qu.:   2.000   3rd Qu.:   2.0
##  Max.   :1.000000   Max.   :3890.000   Max.   :6706.0
##  Flow.Duration       Flow.Packets.s       Flow.Bytes.s
##  Min.   :        1   Min.   :      0.1   Min.   :0.000e+00
##  1st Qu.:        2   1st Qu.:    114.1   1st Qu.:3.137e+03
##  Median :      109   Median :  36036.0   Median :2.124e+05
##  Mean   : 10230750   Mean   : 562956.5   Mean   :6.165e+06
##  3rd Qu.:    44839   3rd Qu.:1000000.0   3rd Qu.:6.000e+06
##  Max.   :119991155   Max.   :3000000.0   Max.   :1.559e+09
##  Fwd.Packet.Length.Mean Bwd.Packet.Length.Mean  Bwd.IAT.Mean
##  Min.   :   0.00        Min.   :   0.00        Min.   :       0
##  1st Qu.:   6.00        1st Qu.:   0.00        1st Qu.:       0
##  Median :   6.00        Median :   6.00        Median :       1
##  Mean   :  24.58        Mean   :  54.50        Mean   :  414913
##  3rd Qu.:  30.75        3rd Qu.:  10.38        3rd Qu.:       3
##  Max.   :1797.62        Max.   :1792.35        Max.   :34909930
##  ACK.Flag.Count    Active.Mean       Label       Inbound
##  Min.   :0.0000   Min.   :       0   0:4949   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:       0   1:4655   1st Qu.:0.0000
##  Median :1.0000   Median :       0            Median :1.0000
##  Mean   :0.5886   Mean   :   68588            Mean   :0.5939
##  3rd Qu.:1.0000   3rd Qu.:       0            3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :10073804            Max.   :1.0000
```

EDA

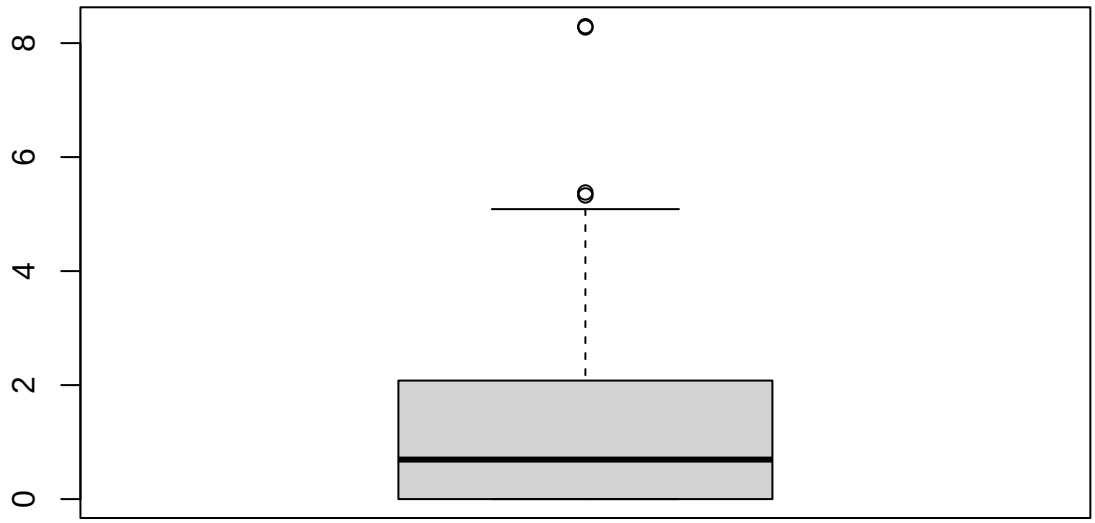
```r
# Create a table with flag count
flag_count_table <- table(syn_selected_dt$SYN.Flag.Count)
flag_count_table
```

```
##
##    0    1
## 9572   32
```

```r
# Create a table w;ith Total BWD Packets
Total_Backward_Packets_table <- table(syn_selected_dt$Total.Backward.Packets)
Total_Backward_Packets_table
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 4011  206 3924  115  217   18   92   11  162    9   57   18   55   22   47   22
##   16   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31
##   22    2   25   11   20   11   16    6   38    7   25   13   47   10   32    6
##   32   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47
##   31    8   21    5   13    7   13    3   19    3    3    9   12    8   10    3
##   48   49   50   51   52   53   54   55   56   57   58   59   60   61   62   63
##    4    2    2    1    2    3    2    1    2    4    8    2    6    1    2    1
##   64   65   67   68   69   70   71   72   74   75   79   80   82   83   84   85
##    1    1    4    2    1    5    2    4    4    2    1    1    2    2    1    1
##   88   92   93   98  101  102  104  106  107  108  110  111  112  114  120  128
##    1    1    1    1    1    2    1    1    1    1    2    2    1    2    1    3
##  129  140  145  146  149  151  154  158  160  172  173  177  178  180  181  186
##    1    1    1    1    2    2    1    1    1    1    1    1    2    1    2    1
##  188  197  203  206  214  215  216  220  222  223  229  232  233  238  240  244
##    1    1    1    1    1    1    1    1    1    2    1    1    1    1    1    1
##  249  250  271  278  282  284  290  298  336  338  345  380  413  420  457  821
##    1    2    1    1    1    2    1    1    1    1    1    1    1    1    1    1
## 1607 1648 1662 2863 6706
##    1    1    1    1    1
```

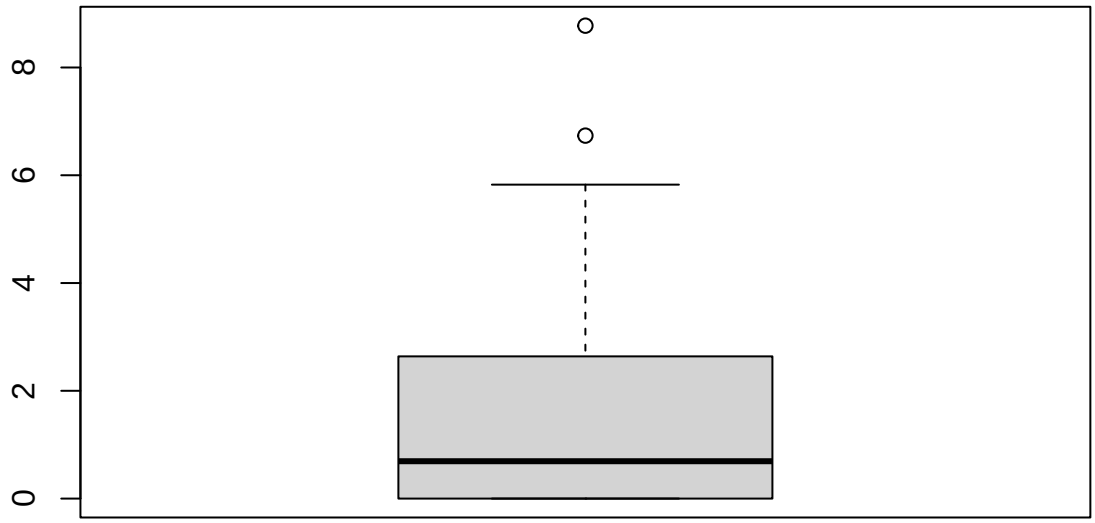
```
boxplot(log(Total_Backward_Packets_table))
```



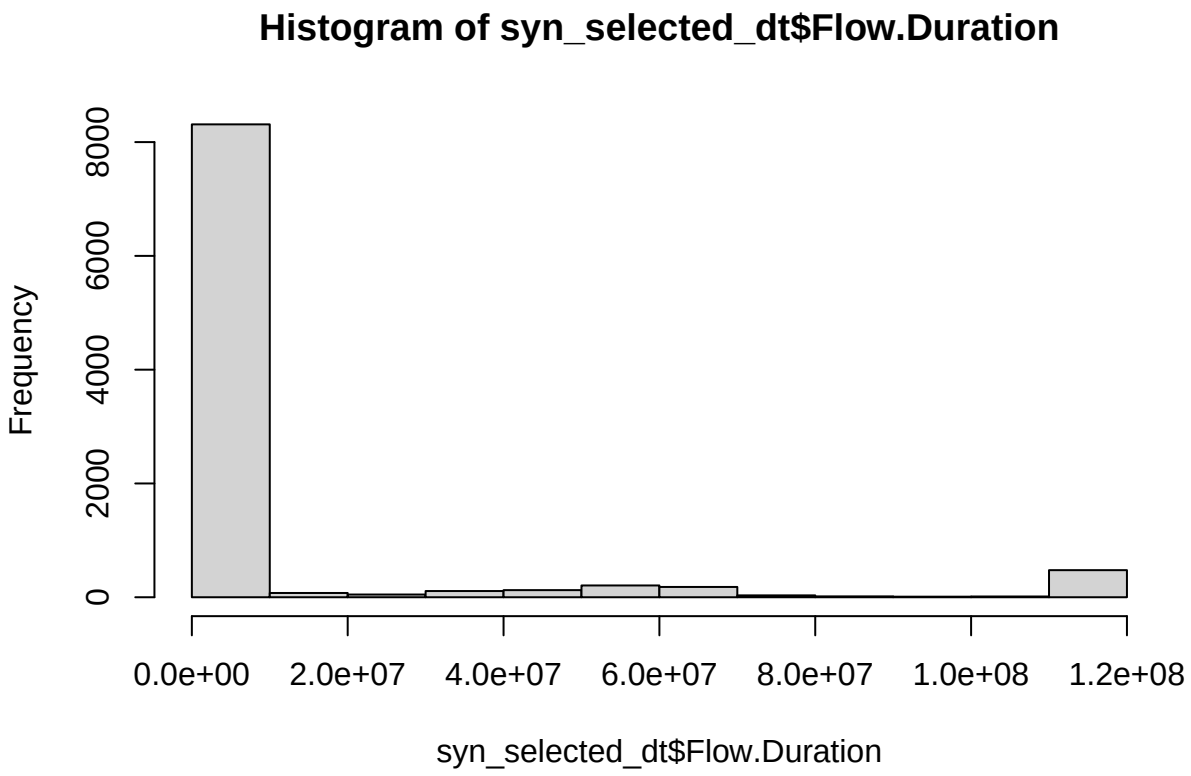

```
# Create a table with Total FWD Packet
Total_Fwd_Packet_table <- table(syn_selected_dt$Total.Fwd.Packets)
Total_Fwd_Packet_table
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
##  842 6476  255  339  103  162   25  158    4  170   11  121    7  100   13   89
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
##   16   62    5   22    7   30    2   31    5   19    6   27    7   31   29   41
##   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
##   14   18   15   23    3   26   24   26    5   22    2   18    4    8    3    8
##   50   51   52   54   56   57   58   59   60   61   62   63   64   65   66   68
##   10    3   34    1    8    1    2    2    7    1    3    2    5    1    1    3
##   69   70   72   74   75   76   77   78   80   82   84   88   89   90   91   94
##    2    2    1    4    1    1    1    1    1    1    2    3    1    2    1    2
##   96   97   98  100  104  107  108  114  116  124  125  126  127  128  129  130
##    2    1    1    1    2    1    2    4    2    3    1    1    1    1    1    5
##  132  134  136  144  148  150  155  158  160  162  164  168  176  184  186  188
##    1    1    1    1    1    1    1    1    2    1    1    1    2    1    1    1
##  224  250  254  274  298  314  380  460  866  893 1000 1639 3890
##    1    1    1    1    1    1    1    1    1    1    1    1    1
```

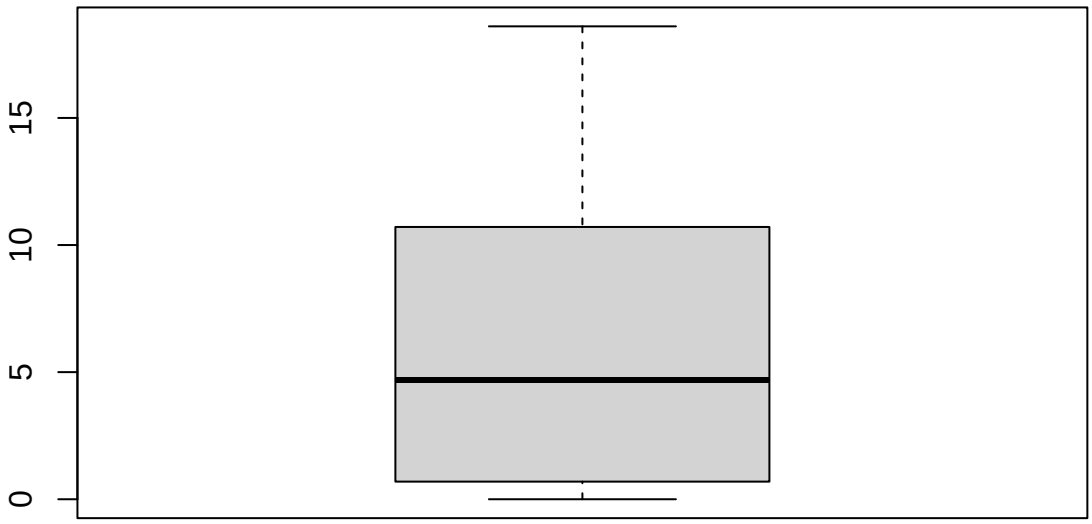
```r
boxplot(log(Total_Fwd_Packet_table))
```




```r
# Create a table with Flow Duration
#Flow_Duration_table <- table(syn_selected_dt$Flow.Duration)
#Flow_Duration_table

hist(syn_selected_dt$Flow.Duration)
```



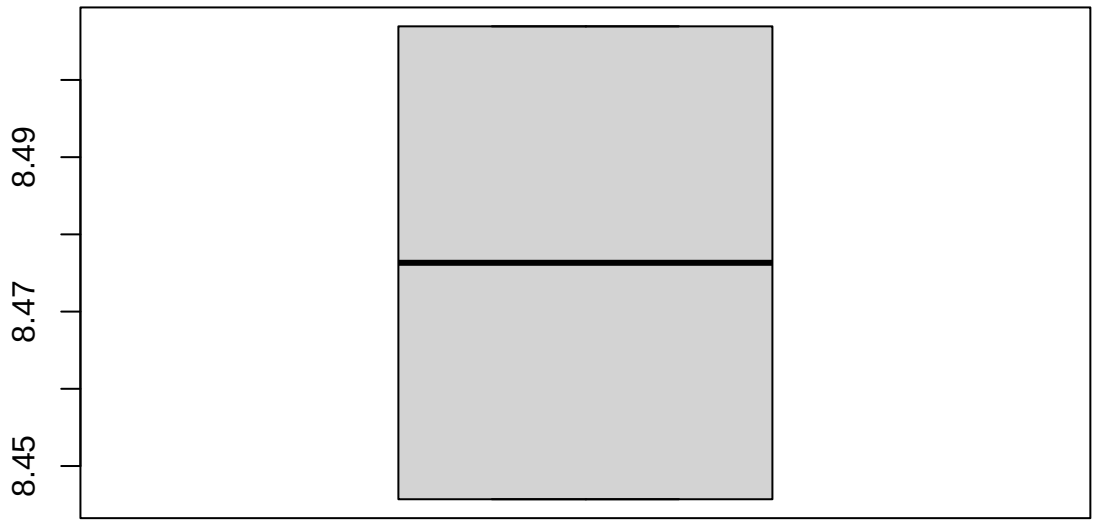

```
boxplot(log(syn_selected_dt$Flow.Duration))
```



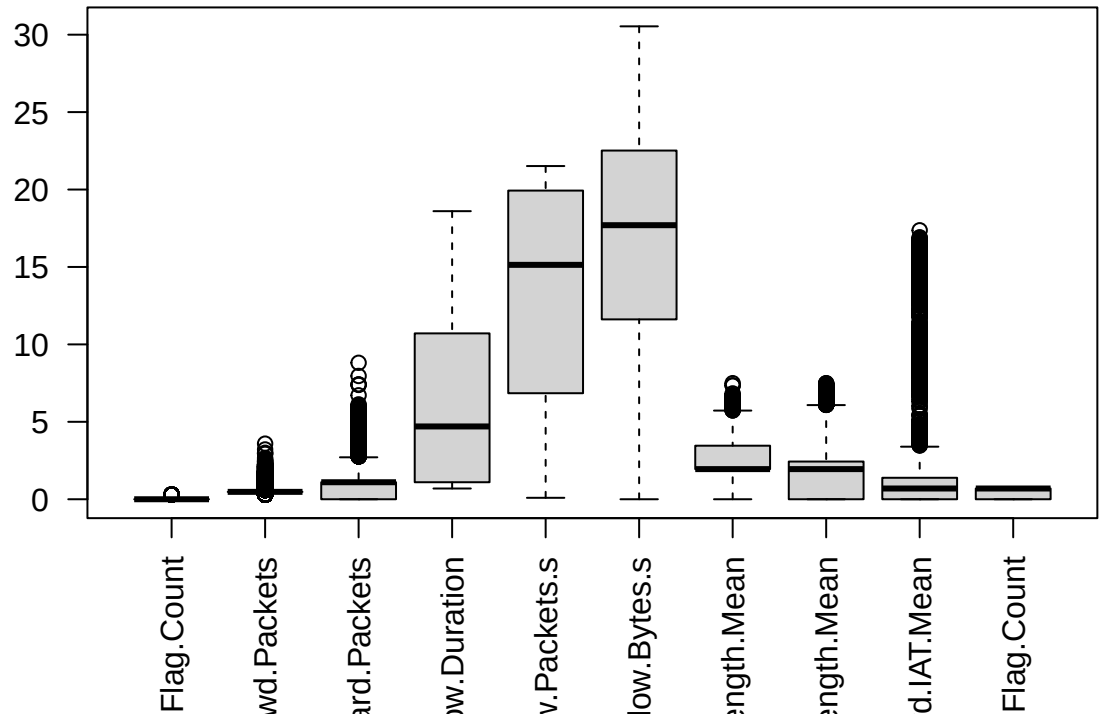

```
# Create a table with Flow labels
Labels_table <- table(syn_selected_dt$Label)
Labels_table
```

```
##
##    0    1
## 4949 4655
```

```
boxplot(log(Labels_table))
```




```r
boxplot(
  log10(syn_selected_dt$SYN.Flag.Count + 1),
  log10(syn_selected_dt$Total.Fwd.Packets + 1),
  log(syn_selected_dt$Total.Backward.Packets + 1),
  log(syn_selected_dt$Flow.Duration + 1),
  log2(syn_selected_dt$Flow.Packets.s + 1),
  log2(syn_selected_dt$Flow.Bytes.s + 1),
  log(syn_selected_dt$Fwd.Packet.Length.Mean + 1),
  log(syn_selected_dt$Bwd.Packet.Length.Mean + 1),
  log(syn_selected_dt$Bwd.IAT.Mean + 1),
  log(syn_selected_dt$ACK.Flag.Count + 1),
  names = c(
    "SYN.Flag.Count",
    "Total.Fwd.Packets",
    "Total.Backward.Packets",
    "Flow.Duration",
    "Flow.Packets.s",
    "Flow.Bytes.s",
    "Fwd.Packet.Length.Mean",
    "Bwd.Packet.Length.Mean",
    "Bwd.IAT.Mean",
    "ACK.Flag.Count"
  ),
  main = "Combined Boxplots",
  las  = 2
)
```




```
t.test(SYN.Flag.Count ~ Label, data = syn_selected_dt)
```

```
##
##  Welch Two Sample t-test
##
## data:  SYN.Flag.Count by Label
## t = 5.2969, df = 5310.1, p-value = 1.225e-07
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.003810280 0.008287857
## sample estimates:
## mean in group 0 mean in group 1
##    0.0062638917    0.0002148228
```

```
t.test(Total.Fwd.Packets ~ Label, data = syn_selected_dt)
```

```
##
##  Welch Two Sample t-test
##
## data:  Total.Fwd.Packets by Label
## t = 7.9754, df = 4965.4, p-value = 1.871e-15
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  5.773029 9.536216
## sample estimates:
## mean in group 0 mean in group 1
##       10.580723        2.926101
```

```r
t.test(Total.Backward.Packets ~ Label, data = syn_selected_dt)
```

```
##
##  Welch Two Sample t-test
##
## data:  Total.Backward.Packets by Label
## t = 6.3323, df = 4950.9, p-value = 2.63e-10
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   7.148127 13.558959
## sample estimates:
## mean in group 0 mean in group 1
##       11.517882        1.164339
```

```r
t.test(Flow.Duration ~ Label, data = syn_selected_dt)
```

```
##
##  Welch Two Sample t-test
##
## data:  Flow.Duration by Label
## t = 15.335, df = 7050.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  7526927 9733348
## sample estimates:
## mean in group 0 mean in group 1
##        14413725         5783587
```

```r
t.test(Flow.Packets.s ~ Label, data = syn_selected_dt)
```

```
##
##  Welch Two Sample t-test
##
## data:  Flow.Packets.s by Label
## t = -34.078, df = 8010.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -596052.5 -531208.7
## sample estimates:
## mean in group 0 mean in group 1
##        289768.2        853398.8
```

```r
t.test(Flow.Bytes.s ~ Label, data = syn_selected_dt)
```

```
##
##  Welch Two Sample t-test
##
## data:  Flow.Bytes.s by Label
## t = 1.2819, df = 5809.3, p-value = 0.1999
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -527220.8 2519290.4
## sample estimates:
## mean in group 0 mean in group 1
##         6647971         5651936
```

```r
t.test(Fwd.Packet.Length.Mean ~ Label, data = syn_selected_dt)
```

```
##
##  Welch Two Sample t-test
##
## data:  Fwd.Packet.Length.Mean by Label
## t = 29.74, df = 5490.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  32.39751 36.97001
## sample estimates:
## mean in group 0 mean in group 1
##       41.386426        6.702665
```

```r
t.test(Bwd.Packet.Length.Mean ~ Label, data = syn_selected_dt)
```

```
##
##  Welch Two Sample t-test
##
## data:  Bwd.Packet.Length.Mean by Label
## t = 31.184, df = 4950.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   94.23263 106.87562
## sample estimates:
## mean in group 0 mean in group 1
##      103.234981        2.680857
```

```r
t.test(Bwd.IAT.Mean ~ Label, data = syn_selected_dt)
```

```
##
##  Welch Two Sample t-test
##
## data:  Bwd.IAT.Mean by Label
## t = -0.76139, df = 8400.6, p-value = 0.4464
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -93397.67  41140.88
## sample estimates:
## mean in group 0 mean in group 1
##        402248.7        428377.1
```

```r
t.test(ACK.Flag.Count ~ Label, data = syn_selected_dt)
```

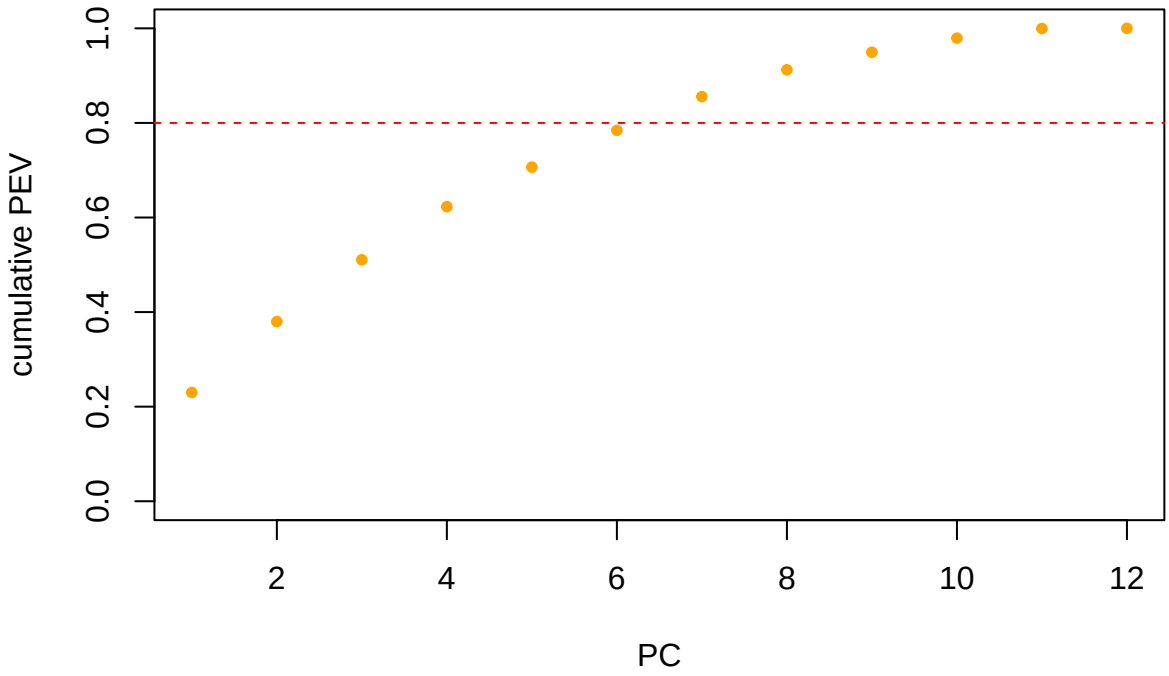
```
##
##  Welch Two Sample t-test
##
## data:  ACK.Flag.Count by Label
## t = -137.53, df = 5100.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.8050720 -0.7824428
## sample estimates:
## mean in group 0 mean in group 1
##       0.2038796       0.9976369
```

PCA

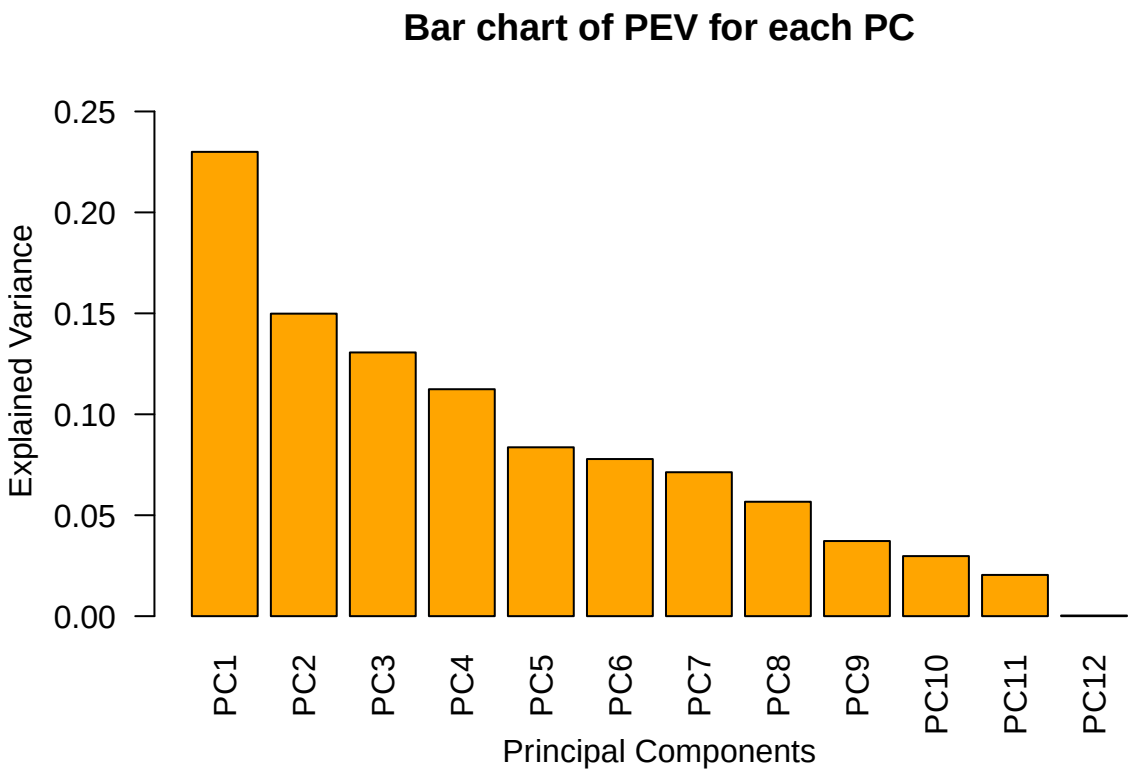
```
#names(syn_selected_dt)
#str(syn_selected_dt)

# Create the groups
all_dt_ex_lable <- as.data.frame(syn_selected_dt[,-12])
```

```
# Calculate PCA on whole dataset

PCA_results <- prcomp(all_dt_ex_lable , center = T, scale. = T)
```

```
# visualising pca results

### 4.1 calculate the proportion of exaplained variance (PEV) from the std values
PCA_results_var <- PCA_results$sdev^2
PCA_result_PEV <- PCA_results_var / sum(PCA_results_var)
### 4.2 plot the cumulative PEV
opar <- par(no.readonly = TRUE)
plot(
  cumsum(PCA_result_PEV),
  ylim = c(0,1),
  xlab = 'PC',
  ylab = 'cumulative PEV',
  pch = 20,
  col = 'orange'
)
abline(h = 0.8, col = 'red', lty = 'dashed')
```




```
par(opar)

### 4.2b barplot of individual PEV (scree plot)
barplot(
  PCA_result_PEV,
  names.arg = paste0("PC", seq_along(PCA_result_PEV)),
  las = 2,
  col = "orange",
  ylab = "Explained Variance",
  xlab = "Principal Components",
  main = "Bar chart of PEV for each PC",
  ylim = c(0, max(PCA_result_PEV) * 1.1)
)
```

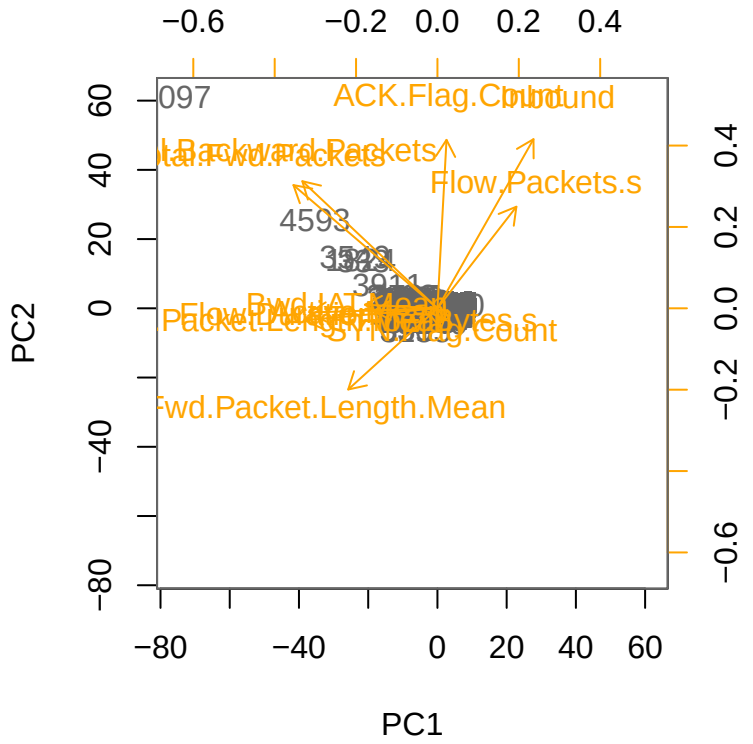



```
### 4.3 get and inspect the loadings
PCA_result_loadings <- PCA_results$rotation
PCA_result_loadings
```

```
##                                PC1          PC2         PC3         PC4
## SYN.Flag.Count           0.01088147 -0.061066837  0.08998625 -0.04360456
## Total.Fwd.Packets       -0.44235661  0.378546539  0.32002481 -0.05716909
## Total.Backward.Packets -0.41499883  0.390475651  0.35979583 -0.06755806
## Flow.Duration           -0.39336880 -0.012866771 -0.45327149  0.01981303
## Flow.Packets.s           0.24257128  0.312145443  0.11926207  0.35496173
## Flow.Bytes.s             0.02385249 -0.036623843  0.11907723  0.73879937
## Fwd.Packet.Length.Mean -0.27414534 -0.249388573 -0.04093110  0.53928740
## Bwd.Packet.Length.Mean -0.40847945 -0.036707468  0.05053877  0.04443462
## Bwd.IAT.Mean            -0.22276644  0.011084863 -0.55742398 -0.02822918
## ACK.Flag.Count           0.02786036  0.517521100 -0.38929716  0.11605330
## Active.Mean             -0.19572474 -0.008101769 -0.15796120  0.07343690
## Inbound                  0.29514239  0.519616962 -0.18633476  0.07475143
##                                PC5         PC6         PC7         PC8
## SYN.Flag.Count          -0.952763963 -0.09553790  0.25039093  0.08032054
## Total.Fwd.Packets       -0.038200380  0.04748078 -0.16707891  0.07874889
## Total.Backward.Packets -0.033340086  0.04718299 -0.17109143  0.10108121
## Flow.Duration           -0.103123107  0.10843992 -0.12053511 -0.16141958
## Flow.Packets.s          -0.118404276  0.01177214 -0.10424964 -0.78980456
## Flow.Bytes.s            -0.075363255  0.13228005 -0.23836477  0.22589324
## Fwd.Packet.Length.Mean  0.066917817 -0.03320398  0.24019469  0.20668578
## Bwd.Packet.Length.Mean  0.126126386  0.06357423  0.61094945 -0.28535572
```

```
## Bwd.IAT.Mean           -0.188424803  0.31359751 -0.36221826 -0.06105527
## ACK.Flag.Count          0.057444358 -0.06366919  0.43146686  0.06905511
## Active.Mean             0.009232037 -0.92053915 -0.21561237 -0.06688831
## Inbound                -0.007847426 -0.04541328  0.05563405  0.37301104
##                                  PC9         PC10        PC11          PC12
## SYN.Flag.Count          0.0008600166  0.003828596  0.01022040  0.0001796424
## Total.Fwd.Packets       0.0889455181  0.014421265  0.02798612 -0.7131661952
## Total.Backward.Packets  0.0697369464 -0.065761510  0.05560521  0.6972663551
## Flow.Duration           0.0716882905  0.714661825 -0.23571457  0.0694777212
## Flow.Packets.s          0.2068236033 -0.060993344 -0.09031994  0.0056438381
## Flow.Bytes.s           -0.5207887951  0.141455038  0.12577598 -0.0035498159
## Fwd.Packet.Length.Mean  0.6455368316 -0.193136472 -0.10020314  0.0095002973
## Bwd.Packet.Length.Mean -0.4731756174 -0.198708926 -0.29874784  0.0010803295
## Bwd.IAT.Mean           -0.0848274947 -0.600536941  0.04322334 -0.0157529148
## ACK.Flag.Count          0.0148902494  0.077775885  0.60151332 -0.0003171667
## Active.Mean            -0.1408199166 -0.112581980 -0.01830221  0.0011198568
## Inbound                -0.0150961042 -0.069470934 -0.67317850 -0.0010179942
```

```r
### 4.4 generate a biplot for PC1 and PC2
opar <- par(no.readonly = TRUE)
biplot(
  PCA_results,
  scale = 0,
  col = c('grey40','orange')
)
```



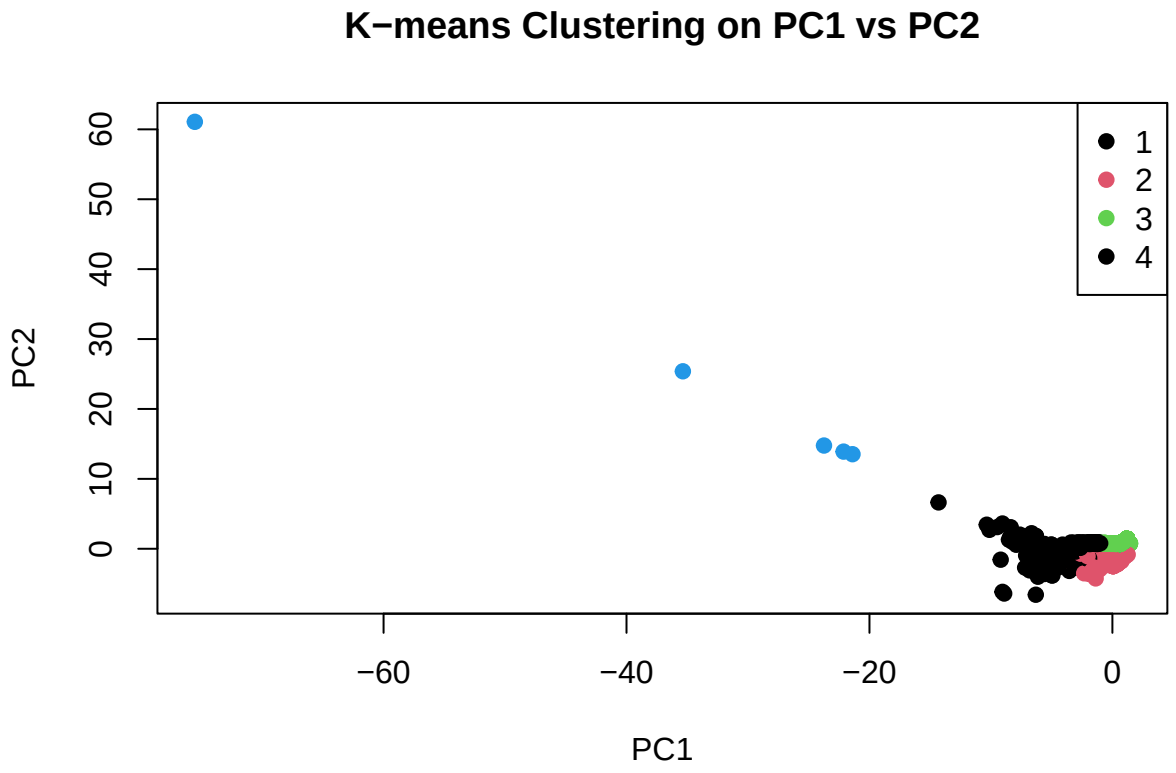
```r
par(opar)
```

so we know its 1 to 6 that we need so we assign only 1 to 6 to the dataframe as a subset of the PCA, then we do Cluster analysis on that. Centres is the amount of clusters, 25 is how many times k means is run. we convert the matrix to a dataframe and then we add the cluster result through "as factor" to the dataframe.

```r
first_six_PC <- PCA_results$x[, 1:6]

kmeans_result <- kmeans(first_six_PC, centers = 4, nstart = 25)
PCA_subset <- as.data.frame(first_six_PC)
PCA_subset$cluster <- as.factor(kmeans_result$cluster)
```

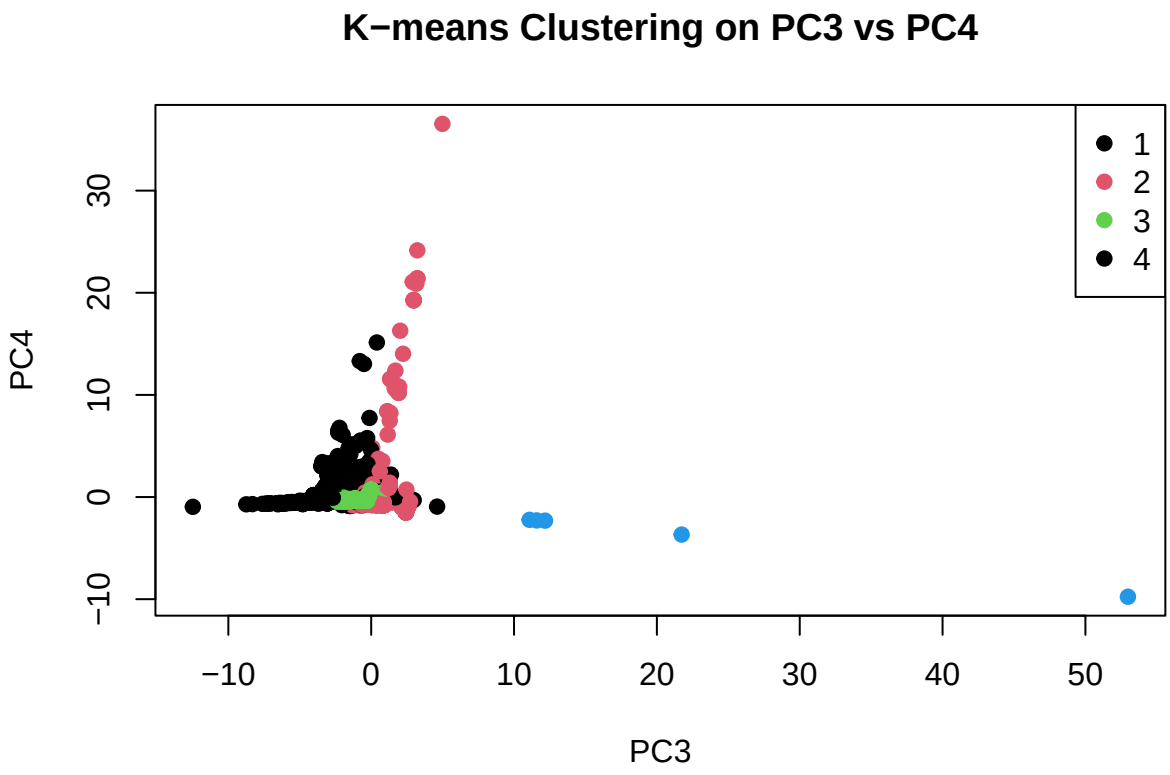
You can only plot 2 PCs and see what clusters they form against eachother to visualise it so here are two examples of that.

```r
plot(
  PCA_subset$PC1, PCA_subset$PC2,
  col = PCA_subset$cluster,
  pch = 19,
  xlab = "PC1", ylab = "PC2",
  main = "K-means Clustering on PC1 vs PC2"
)
legend("topright", legend = levels(PCA_subset$cluster), col = 1:3, pch = 19)
```




```r
plot(
  PCA_subset$PC3, PCA_subset$PC4,
  col = PCA_subset$cluster,
```

```
  pch = 19,
  xlab = "PC3", ylab = "PC4",
  main = "K-means Clustering on PC3 vs PC4"
)
legend("topright", legend = levels(PCA_subset$cluster), col = 1:3, pch = 19)
```



Save slected dataset

```
# Save the final cleaned dataset
write.csv(syn_selected_dt , "syn seleceted dt.csv", row.names = FALSE)
```